1-27-2020

Quantifying the association between discrete event time series with applications to digital forensics

Christopher Galbraith
University of California, Irvine

Padhraic Smyth
University of California, Irvine

Hal S. Stern
University of California, Irvine

Quantifying the association between discrete event time series with applications to digital forensics

Abstract

We consider the problem of quantifying the degree of association between pairs of discrete event time series, with potential applications in forensic and cybersecurity settings. We focus in particular on the case where two associated event series exhibit temporal clustering such that the occurrence of one type of event at a particular time increases the likelihood that an event of the other type will also occur nearby in time. We pursue a non-parametric approach to the problem and investigate various score functions to quantify association, including characteristics of marked point processes and summary statistics of interevent times. Two techniques are proposed for assessing the significance of the measured degree of association: a population-based approach to calculating score-based likelihood ratios when a sample from a relevant population is available, and a resampling approach to computing coincidental match probabilities when only a single pair of event series is available. The methods are applied to simulated data and to two real world data sets consisting of logs of computer activity and achieve accurate results across all data sets.

Keywords

Discrete events, Forensics, Likelihood ratio, Spatial statistics, Time series

Disciplines

Forensic Science and Technology

Comments

This article is published as Galbraith, C., Smyth, P. and Stern, H.S. (2020), Quantifying the association between discrete event time series with applications to digital forensics. *Journal of the Royal Statistical Society Series A.* doi:10.1111/rssa.12549. Posted with permission from CSAFE.

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2020)

Quantifying the association between discrete event time series with applications to digital forensics

Christopher Galbraith, Padhraic Smyth and Hal S. Stern

University of California, Irvine, USA

[Received February 2019. Revised November 2019]

Summary. We consider the problem of quantifying the degree of association between pairs of discrete event time series, with potential applications in forensic and cybersecurity settings. We focus in particular on the case where two associated event series exhibit temporal clustering such that the occurrence of one type of event at a particular time increases the likelihood that an event of the other type will also occur nearby in time. We pursue a non-parametric approach to the problem and investigate various score functions to quantify association, including characteristics of marked point processes and summary statistics of interevent times. Two techniques are proposed for assessing the significance of the measured degree of association: a population-based approach to calculating score-based likelihood ratios when a sample from a relevant population is available, and a resampling approach to computing coincidental match probabilities when only a single pair of event series is available. The methods are applied to simulated data and to two real world data sets consisting of logs of computer activity and achieve accurate results across all data sets.

Keywords: Discrete events; Forensics; Likelihood ratio; Spatial statistics; Time series

1. Introduction

Forensic analysis involves analysing observed evidence during a legal investigation. This can be in the context of civil or criminal investigations. For the present study we focus on forensic analysis in criminal settings. Statistical techniques have played a key role in forensic analysis, providing forensic investigators with tools that enable them to make robust inferences from limited and noisy data. The best-known example in this context is the use of likelihood ratio techniques for assessing the strength of the evidence that a deoxyribonucleic acid (DNA) sample from a crime scene is a match to a suspect's DNA sample (Evett and Weir, 1998; Myers *et al.*, 2011). For other types of evidence, such as fingerprints, shoeprints, bullet casing impressions and glass fragments, the development of quantitative methodologies (such as likelihood ratio techniques) is more challenging (Stern, 2017). In particular, there are significant challenges in developing realistic statistical models, both for capturing the process by which the evidential data are produced and for modelling the inherent variability of such data from a relevant population.

In this context, the increased prevalence of *digital evidence* presents both opportunities and challenges from a statistical perspective. Digital evidence is typically defined as evidence that is obtained from a digital device, such as a mobile phone or a computer, where the evidence is associated with a crime scene or with a suspect. As the use of digital devices has increased, so also has the amount of user-generated event data collected by these devices. Such data can be obtained

Address for correspondence: Christopher Galbraith, Department of Statistics, University of California, Irvine, Bren Hall 2019, Irvine, CA 92697, USA.
E-mail: galbraic@uci.edu

from logs of time-stamped events stored either directly on a device such as a mobile phone or computer or stored on a user's account in the cloud (Oh *et al.*, 2011; Roussev and McCulley, 2016). Examples of such events include user actions when using particular software, searching or browsing activities in a web browser and communicating via e-mail or text messaging. This type of user-generated event data tends to be both

(a) inhomogeneous over time (often with circadian rhythms) and
(b) bursty, with brief periods of high activity followed by periods of no activity (e.g. Radicchi (2009)).

These general characteristics pose some challenges from the perspective of developing appropriate statistical models. Nonetheless, despite the challenges, there is a growing need for quantitative statistical approaches in digital forensics, given that existing forensic tools for digital evidence focus primarily on supporting the process of extraction of information from digital devices followed by exploratory analysis (for example see Casey (2011), Roussev (2016) and Årnes (2017)), with little support for statistical quantification.

In this paper we focus on the problem of quantifying the degree of association between two event time series. As an example, consider the case where one event series A consists of a log of time-stamped events (such as log-ins, file access events, browsing and messaging) generated on a device that is associated with a crime (e.g. on a mobile phone found at a crime scene). A second event series B consists of a log of similar events associated with a suspect (e.g. user-generated events recorded on a device that is owned by the suspect). The evidence consists of both event series A and B and the question of interest is to determine how likely it is that the two series were generated by the same individual.

We develop and evaluate non-parametric methods for quantifying the association between pairs of potentially related discrete event series. We focus on a particular (and common) situation where event dependence arises because one type of event tends to occur within bursts of the other type. We then assess the strength or degree of the association in two scenarios:

(a) using score-based likelihood ratios when multiple pairs of discrete event series are available to serve as reference data and
(b) using a resampling approach to compute the probability of a coincidental match when only a single pair of event series is available.

The remainder of the paper is organized as follows: Section 2 provides a formal problem statement and introduces notation. Section 3 outlines relevant background on common approaches to assessing strength of association in a forensic context and discusses related work from the spatial statistics literature. Section 4 discusses some measures that are used to quantify association between pairs of event series. Section 5 describes two methods to assess the strength of association for a given observed measure. Section 6 presents results of applying the proposed methods on simulated and real world data. Finally, Section 7 provides our concluding remarks.

All techniques that are described below are implemented in the open source R package `assocr` which is available from `https://github.com/UCIDataLab/assocr`. The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/1467985x/series-a-datasets`.

2. Problem statement and notation

Consider a pair of user-generated *event series*, where each event series is defined by a set of

times when events of the appropriate type occurred, e.g. series A and series B with n_A and n_B events of type A and type B respectively. Equivalently, the pair of event series (A, B) can be thought of as a *temporal marked point process* M (e.g. Daley and Vere-Jones (2003)), with $M = (A, B) = \{(t_j, m(t_j))\}$ for $j = 1, \ldots, n = n_A + n_B$ where $t_j \in \mathbb{R}^+$ and $m(t_j) \in \{A, B\}$ are the time and type (or mark) of the jth event respectively. For example, series A and B could consist of time-stamped events corresponding to activity from two user accounts (e.g. accounts on a social media platform such as Twitter) where we may be interested in determining whether the two accounts belong to the same individual.

We focus on the case where the two event series exhibit temporal clustering such that the occurrence of one type of event at a particular time increases the likelihood that an event of the other type will also occur nearby in time. In contrast, we can also have 'negative association', where one type of event tends to repel the other type (e.g. when one individual uses two devices or accounts at distinct and clearly separated times). This alternative is not pursued in this paper, although it may be possible to adapt the present framework to such situations.

Fig. 1 provides an example of the types of pairs of temporally clustered event series that we focus on. The data consist of two pairs of event series of browser actions, (A_i, B_i) and (A_j, B_j) generated by users i and j respectively, taken from the case-study that is discussed later in Section 6.2. From Fig. 1 it is visually apparent that A_i and B_i are associated with each other, as are A_j and B_j.

In this general context we address the problem of developing methods to quantify the likelihood of observing the pair of event series (A, B) under different hypotheses regarding their source. In particular, we focus on two specific aspects of this problem:

(a) investigating suitable measures $\Delta(A, B)$ to quantify the association between two event series A and B, and
(b) quantifying the likelihoods of observing the pair (A, B)—or more precisely the likelihood of observing the relevant summary $\Delta(A, B)$—under the hypotheses that the series were generated by the same source or by different sources. We shall refer to this (second) aspect of the problem as *assessing the strength or degree of association* between the two event series.

We address the first question by leveraging ideas from the marked point process literature where a variety of techniques have been developed for measuring association between marks, particularly for spatial point processes (Illian *et al.*, 2008; Baddeley *et al.*, 2015). Real world pairs of event series $M = (A, B)$ of user-generated event data can exhibit significant burstiness and inhomogeneity over time (e.g. as in Fig. 1), making it challenging to develop robust parametric models of association between A and B. For this reason we pursue non-parametric measures

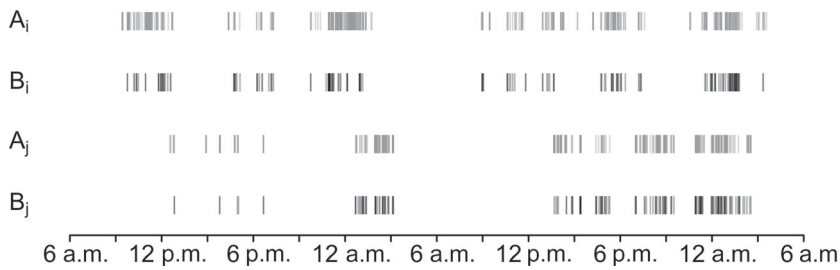

Fig. 1. Example of temporal marked point processes from two individuals (i and j) taken from the case-study of Section 6.2: A and B events generated by the same individual tend to cluster temporally, with less clustering in time for A and B events from different users

of association between temporal processes, particularly measures based on near-neighbour and interevent time characteristics.

To address the second question, the quantification of the likelihood of observing A and B (or more precisely $\Delta(A, B)$) under competing hypotheses about the source(s) of the series, we investigate two methods. The first is a population-based approach where we have realizations from N relevant pairs of processes $M_i = (A_i, B_i)$ for $i = 1, \ldots, N$. The second is a resampling approach when only a single pair M is available, i.e. we do not have access to a sample from a relevant population of realizations.

3. Background on approaches to assessing the strength of association

We shall discuss two general threads of related work in this section:

(a) the use of the likelihood ratio in forensics for the assessment of strength of evidence and
(b) modelling dependence in marked point processes.

3.1. Likelihood ratio methods in forensic science

The *likelihood ratio* is widely accepted in the forensic science community as 'a logically defensible way' to assess the strength of evidence (Willis *et al.*, 2016) having been applied in a variety of forensic disciplines, including handwriting, speech, fingerprints and DNA (Aitken and Stoney, 1991; Evett and Weir, 1998; Champod and Meuwly, 2000; Champod and Evett, 2001; Bozza *et al.*, 2008). (The term 'evidence' typically refers to outcomes of forensic examinations that may be used by legal decision makers in a court of law to reach a belief about a proposition (Willis *et al.*, 2016). In this paper, evidence refers to the observed pair of event time series of interest $M = (A, B)$.) See Stern (2017) for a thorough discussion of the likelihood ratio and its application across a variety of forensic disciplines. The likelihood ratio compares the probability of the evidence under two competing hypotheses, denoted H_s and H_d. Given two items of evidence (A, B), these hypotheses are

$$H_s : (A, B) \text{ came from the same source,}$$
$$H_d : (A, B) \text{ came from different sources.}$$

(In this paper, 'source' refers to either an individual or user account, and 'came from' can be interpreted as 'generated by'. Thus, H_s is the proposition that (A, B) were generated by the same individual or user account, and similarly for H_d.) The likelihood ratio is the quantity that summarizes the information that is contained in the data (A, B). By applying Bayes's theorem, we obtain

$$\underbrace{\frac{\Pr(H_s|A, B, I)}{\Pr(H_d|A, B, I)}}_{a\ posteriori\ \text{odds}} = \overbrace{\frac{f(A, B|H_s, I)}{f(A, B|H_d, I)}}^{\text{likelihood ratio}} \underbrace{\frac{\Pr(H_s|I)}{\Pr(H_d|I)}}_{a\ priori\ \text{odds}} \tag{1}$$

where I is all of the information that is available to the decision maker before the introduction of the evidence (A, B), and f is either a conditional probability mass or density function depending on whether A and B are discrete or continuous respectively. Formula (1) shows that the likelihood ratio arises naturally as the mechanism for updating the prior odds of H_s (relative to H_d) to obtain the posterior odds. We suppress the notation for conditioning on I for the remainder of this paper.

Likelihood ratios require a full generative model for A and B, however, which can be challenging with high dimensional data and may require potentially unrealistic parametric assumptions

depending on the context. An alternative approach that is gaining popularity (e.g. Bolck *et al.* (2015) and Meuwly *et al.* (2017)) is instead to measure similarity between A and B via a *score function* $\Delta(A, B)$ that is usually univariate and continuous. Typically, low scores indicate that the samples are similar, whereas high scores indicate considerable differences. The *score-based likelihood ratio* (SLR) can then be defined as

$$\mathrm{SLR}_{\Delta} = \frac{g\{\Delta(A, B) = \delta | H_{\mathrm{s}}\}}{g\{\Delta(A, B) = \delta | H_{\mathrm{d}}\}} \tag{2}$$

where g denotes the conditional probability density function of $\Delta(A, B)$, and g is typically straightforward to estimate via standard parametric or non-parametric techniques.

The numerator of the SLR in equation (2) can be interpreted as the likelihood of observing the score $\Delta(A, B) = \delta$ if A and B came from the same source. The interpretation of the denominator is the likelihood of observing this score if A and B came from different sources. However, there is ambiguity in the definition of 'different sources' (for example, Hepler *et al.* (2012) provided three possible interpretations of H_{d} that yield different SLRs). In this paper, we focus on the 'general match' interpretation, which holds that the denominator is the likelihood of observing the score $\Delta(A, B) = \delta$ if A is from a randomly selected source from a relevant population and paired with B from a different randomly selected source from a relevant population. This approach is often used in biometrics for example (Ross *et al.*, 2006).

Galbraith and Smyth (2017) investigated the potential of SLRs for assessing the strength of association by using a combination of near-neighbour score functions and population-based inference (see Section 5.1). In the present paper we extend these ideas to a broader and more general framework. In particular, we investigate score functions by using both neighbourhood characteristics and interevent times and extend our approach to use randomization techniques for situations without population data. The resampling approach in particular opens up the proposed methodology to a much broader range of applications in practice, given that it relaxes the need for data from a reference population.

3.2. Event series analysis

Measuring the association between event series is an issue that arises in various application problems. For example, a common problem in spatial statistics is determining the relationship between point patterns (i.e. marked point processes) by performing inference (either analytical or numerical) under a null model that typically assumes some form of independence between the patterns. One of the most well-known techniques is Ripley's cross-K function (Dixon, 2014), which measures the number of occurrences of one type of point within a given radius r of the other type of point as a function of r. Under certain assumptions on the processes themselves (e.g. stationarity) and the relationship between the processes (e.g. complete spatial randomness), significance tests can be performed to determine whether the observed function is consistent with the assumed relationship (e.g. Diggle and Chetwynd (1991) and Gaines *et al.* (2000)).

A popular alternative to the analytical significance test uses *simulation envelope* techniques which compute a summary function of the observed point patterns (such as Ripley's cross-K function) and compare the observed function with the envelope of a set of functions obtained from simulations of the null model (Baddeley *et al.*, 2014). Numerous methods exist for computing simulation envelopes, but the most relevant to the present work use some form of a bootstrap to perform the simulation. Loh (2008) fixed the spatial locations and resampled the marks of the points to obtain confidence envelopes of spatial correlation functions. Niehof and Morley (2012) kept the marks fixed but resampled the (temporal) locations of the points by incorporating a moving block bootstrap (Synowiecki, 2007). Our method is conceptually related

to simulation envelopes obtained by bootstrapping. However, we focus on a single point (e.g. a single score) rather than a function and the focus of our analysis is on data that are both inhomogeneous and bursty.

Another related line of work has been developed in the neuroscience literature. One approach in that context is the spike train reliability statistic R that was introduced by Hunter and Milton (2003) to measure the association of neural spike trains. R is taken to be the mean normalized exponentiated time between events in one series and the corresponding nearest neighbours in another series. This is related to the interevent time score functions that we introduce in Section 4.2. However, no method to assess the statistical significance of R has been provided. Another method in this context is event synchronization, as proposed by Quiroga *et al.* (2002) for measuring correlation in left-hand and right-hand electroencephalography channels. More recently, the technique has also been used in climatological applications (Boers *et al.*, 2016; Malik *et al.*, 2010). Event synchronization is similar to the reliability statistic in that it is a function of the time between events in one series to nearest neighbours in the other, but it also takes into account the marginal interevent times (i.e. the time between events for the subprocess restricted to events of a single type). In this manner, event synchronization is similar to the signal-to-noise ratio (SNR) that we utilize to determine the detectability of association (see Section 6.1).

Building on work from both of the aforementioned domains, Donges *et al.* (2016) proposed a framework called event coincidence analysis with an open source software package for quantifying the strength, directionality and time lag of relationships between event series (Siegmund *et al.*, 2017). Event coincidence analysis focuses on coincidences, which were defined by Donges *et al.* (2016) as the occurrence of at least one event in each series in some (τ-lagged) time window ΔT. Under the assumptions that both series are independent Poisson processes and that events are rare (i.e. the number of events multiplied by ΔT is sufficiently smaller than the period spanned by the series), analytical significance tests for the number of observed coincidences have been derived. Donges *et al.* (2016) relaxed these assumptions by proposing two surrogate-data-based tests that rely on simulating realizations of both processes via either random event times or from some prescribed interevent time distribution. Conceptually this method is the most similar to our proposed technique but requires the specification of both a time lag τ and coincidence window ΔT before performing any analysis. One could perform the significance tests for multiple values of τ and ΔT, but then the analysis would suffer from multiple-testing issues.

4. Measures of association

We investigate several score functions that characterize the association between a pair of event time series. Two types of score function are considered, based on

(a) nearest neighbour characteristics and
(b) summary statistics of interevent times.

Note that the score functions rely on the notion of a *reference point*, which is a term that is related to the Palm distribution (Hanisch, 1984). For practical purposes, we define the reference point as an arbitrarily selected event in the pair of event series $M = (A, B)$. The reference point may be of either type.

4.1. Score functions using nearest neighbours

The *coefficient of segregation* (Pielou, 1977) is a function of the ratio of

(a) the probabilities that a (randomly chosen) reference point and its nearest neighbour have different marks to
(b) the same probability for independent marks, defined as

$$S(A,B) = 1 - \frac{p_{AB} + p_{BA}}{p_A p_{\cdot B} + p_B p_{\cdot A}}. \tag{3}$$

Here p_{AB} (or p_{BA}) is the joint probability that the reference point is type A and its nearest neighbour in time is type B (or vice versa), p_A and p_B are the relative frequencies of the two types of points and $p_{\cdot A}$ (or $p_{\cdot B}$) is the probability that the nearest neighbour is type A (or B) irrespective of the type of the reference point. These probabilities are naturally estimated by the empirical relative frequencies of the appropriate events as observed in the data:

$$\begin{aligned} \hat{p}_A &= \frac{n_A}{n_A + n_B} = \frac{n_A}{n}, \\ \hat{p}_{AB} &= \frac{1}{n_A} \sum_{j=1}^{n} \mathbb{I}\{m(t_j) = A\} \mathbb{I}[m\{z_1(t_j)\} = B], \\ \hat{p}_{\cdot B} &= \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}[m\{z_1(t_j)\} = B] \end{aligned} \tag{4}$$

where $z_1(t_j)$ denotes the nearest neighbour of the point t_j, $m(\cdot)$ the mark of the given point and $\mathbb{I}(\cdot)$ the indicator function. Similar definitions hold for $\hat{p}_B$, $\hat{p}_{BA}$ and $\hat{p}_{\cdot A}$.

Note that $S(A,B) \in [-1,1]$. If the reference point and its nearest neighbour always are the same type, then $p_{AB} = p_{BA} = 0$ and $S(A,B) = 1$. This corresponds to repulsion or segregation of points by their mark (i.e. points of type A always occur near each other and never near points of type B and vice versa). If the reference point and its nearest neighbour always have different marks, then $p_{AA} = p_{BB} = 0$, which implies that $p_{\cdot A} = p_{BA}$ and $p_{\cdot B} = p_{AB}$, so $S(A,B) < 0$ with a minimum of $S(A,B) = -1$ if $p_A = p_B = \frac{1}{2}$. This is the opposite of segregation, indicating that points of different marks are attracted to one another. If the marks are independent then $S(A,B)$ will tend to 0 as the size of the observed data set grows since $p_{AB} \approx p_A p_{\cdot B}$ and $p_{BA} \approx p_B p_{\cdot A}$.

4.2. Score functions using interevent times

We also investigate score functions based on interevent times for a pair of event time series. In principle, we expect the interevent times to carry more information than the nearest neighbour characteristics that are used in the coefficient of segregation. We construct distributions of interevent times by fixing the events from one series (say B) and measuring the time from each event in B to the closest event in the other series A. In this paper we define 'closest in time' to mean the closest event either forwards or backwards in time, but a directional definition (e.g. forwards or backwards only) could also be used. In practice we define the series with fewer events as B and measure the interevent times to series A (thus $n_B < n_A$).

Let $\mathcal{T}_{BA}$ represent the set of n_B interevent times from B to A and define it as follows:

$$\mathcal{T}_{BA} \equiv \{\tau_{BA,j} : j = 1, \ldots, n_B\} \qquad \tau_{BA,j} = \min_{k \in \{1,\ldots,n_A\}} |t_{b,j} - t_{a,k}| \tag{5}$$

and $t_{b,j}$ denotes the jth event time of series B, and similarly for series A. See Fig. 2 for an illustration. If events of type B are clustered in time with events of type A, then the interevent times $\mathcal{T}_{BA}$ tend to be smaller than if A and B events are generated independently. A variety of characteristics of the distribution of interevent times could be used as score functions. In this paper we consider the mean interevent time from B to A,

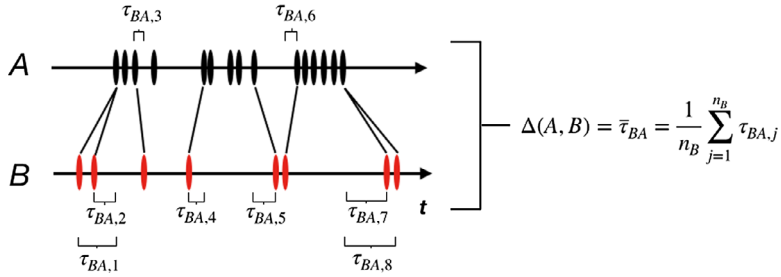

Fig. 2. Example mean interevent time calculation

$$\overline{\mathcal{T}}_{BA} = \frac{1}{n_B}\sum_{j=1}^{n_B} \tau_{BA,j}, \tag{6}$$

and the median interevent time from B to A, med($\mathcal{T}_{BA}$).

5. Assessing the degree of association

Suppose that we are given a pair of event time series $M^* = (A^*, B^*)$ and a particular score function Δ, such as one of those defined in the previous section. We wish to assess the degree of association between A^* and B^*. To do so, we must consider the likelihood of observing $\Delta(A^*, B^*)$ under two competing hypotheses, namely that A^* and B^* were generated by the same source, or that they were generated by two different sources. We investigate two different methods in this context:

(a) a population-based approach in which we have realizations from N pairs of processes, and
(b) a resampling approach when only a single pair M^* is available.

5.1. Population-based approach

We begin by considering a situation in which we have a sample of N pairs of event time series $M_i = (A_i, B_i)$ for $i = 1, \ldots, N$ from a relevant reference population. We consider the case where both event series A_i and B_i in each pair are from the same source, and that each pair M_i is from a different individual i. Let each pair M_i have $n_i = n_{i,A} + n_{i,B}$ events, where $n_{i,A}$ (or $n_{i,B}$) denotes the ith individual's number of events of type A (or B). In a forensic setting, this sample could correspond to a database of event series from potential suspects in an investigation. We use this sample to estimate score-based likelihood ratios according to equation (2).

To compute the SLR for a particular pair of event series (A^*, B^*), we construct empirical estimates of the conditional densities of $\Delta(A, B)$ given the two hypotheses H_s and H_d, using the sample of N pairs. We construct a reference data set of all N^2 pairwise combinations of series, denoted $\mathcal{D} \equiv \{(A_j, B_k) : j, k \in \{1, \ldots, N\}\}$, where each of the N series of type B are paired up with each of the N series of type A. Given a new pair of series (A^*, B^*) that is not from $\mathcal{D}$, we can use the scores of all of the N same-source pairs, $\mathcal{D}_s = \{(A_j, B_j) : j = 1, \ldots, N\}$, to estimate the probability density function in the numerator of equation (2) and the scores of all $N^2 - N$ pairs with different sources, $\mathcal{D}_d = \{(A_j, B_k) : j, k \in \{1, \ldots, N\}, j \neq k\}$, to estimate the probability density function in the denominator. Once we have obtained the estimated probability density functions $\hat{g}\{\Delta(A, B)|H_s\}$ and $\hat{g}\{\Delta(A, B)|H_d\}$ we evaluate their ratio at $\Delta(A^*, B^*) = \delta^*$ to obtain $\widehat{\text{SLR}}_\Delta$.

To illustrate this approach, consider the score function for the mean interevent time proposed in equation (6) and let $\Delta(A, B) = \overline{\mathcal{T}}_{BA}$ for any given pair of event time series (A, B). Thus,

$\delta^* = \overline{T}_{B^*A^*}$ is the observed value that we would like to estimate the SLR for via

$$\widehat{\mathrm{SLR}}_{\overline{T}_{BA}} = \frac{\hat{g}[\overline{T}_{BA} = \overline{T}_{B^*A^*} | \{\overline{T}_{BA} : (A, B) \in \mathcal{D}_\mathrm{s}\}]}{\hat{g}[\overline{T}_{BA} = \overline{T}_{B^*A^*} | \{\overline{T}_{BA} : (A, B) \in \mathcal{D}_\mathrm{d}\}]}. \tag{7}$$

To evaluate the out-of-sample performance of this population-based approach we use *leave-pairs-out cross-validation* to estimate the SLR for every pairwise combination that is available in $\mathcal{D}$. Let $(A^*, B^*) = (A_l, B_m)$ be an arbitrary pair from $\mathcal{D}$, where l and m may or may not be equal. Given (A_l, B_m) let $\mathcal{D}_\mathrm{s}^* = \{(A_j, B_j) : j \in \{1, \ldots, N\} \setminus \{l, m\}\}$ and $\mathcal{D}_\mathrm{d}^* = \{(A_j, B_k) : j, k \in \{1, \ldots, N\} \setminus \{l, m\}, j \neq k\}$ be the sets that are used to compute the scores for estimating the probability density functions of the numerator and denominator of equation (2) respectively for (A^*, B^*). To estimate these densities, we use a kernel density estimator with a Gaussian kernel and a computationally simple rule-of-thumb bandwidth selector (Scott, 1992), given that we need to estimate $O(N^2)$ different kernel densities to perform cross-validation. (The scores that are considered are bounded, and an unconstrained kernel density estimator will push probability mass outside these bounds (e.g. below 0 for interevent time score functions). More sophisticated methods could be used to estimate these densities, but for simplicity and computational efficiency we used a generic kernel density estimator method.)

5.2. Resampling approach

The population-based approach above is useful when a reference population of pairs of event series is available, e.g. user-generated data from a relevant population of users. However, there are many situations in practice where data from a population of users is not readily available. Furthermore, even when a population is available, it is often quite difficult to define the *relevant* reference population of interest in a forensic setting. Should the relevant population be a sample from all individuals in general, or from everyone who matches the description of a suspect in a given region, or from some other group? (See Stern (2017) for additional discussion of this issue.) To address these potential problems we propose below a resampling approach that computes coincidental match probabilities (CMPs) by using only a single pair of event series.

5.2.1. Coincidental match probability

We define the CMP as the probability that two series A^* and B^* exhibit the characteristics of a same-source pair, i.e. a small value of $\Delta(A^*, B^*)$, by chance given that they are from different sources. CMPs are intrinsically related to the denominator of the likelihood ratio, i.e. the conditional likelihood of observing the value $\Delta(A^*, B^*)$ given that the series have different sources. Further, CMPs share conceptual similarities with random-match probabilities that are frequently used in forensics, particularly in DNA analysis (Thompson and Newman, 2015). When computing random-match probabilities, forensic scientists first determine whether two samples match, and, if so, they then compute the probability that the samples match by chance. When calculating CMPs, however, we do not first attempt to determine whether the series A^* and B^* are from the same source but instead calculate the probability that they exhibit the observed degree of association by chance. Thus, CMPs and random-match probabilities are related but have different interpretations.

To estimate the CMP we use resampling in time to simulate new realizations of the event series A^* under a null model for different-source data and thus induce a distribution of scores under this model. Specifically, given an observed pair (A^*, B^*) we define the CMP as the probability that a randomly sampled pair (A', B^*) under the different-source model has a more extreme

score $\Delta(A', B^*)$ (indicating greater similarity) than the observed score $\Delta(A^*, B^*) = \delta^*$:

$$\mathrm{CMP}_\Delta = \Pr\{\Delta(A', B^*) < \delta^* | H_\mathrm{d}\}. \tag{8}$$

('More extreme' here is related to the notion of hypothesis tests and can be defined as either one or two sided. The definition of CMP in equation (8) assumes a one-sided test where the observed score for same-source pairs tends to be less than that of different-source pairs.) We propose the following natural estimator for the CMP:

$$\widehat{\mathrm{CMP}}_\Delta = \frac{1}{n_\mathrm{sim}} \sum_{l=1}^{n_\mathrm{sim}} \mathbb{I}\{\Delta(A^{(l)}, B^*) < \delta^*\} \tag{9}$$

where $A^{(l)}$ for $l = 1, \ldots, n_\mathrm{sim}$ are randomly sampled under H_d by using the null model for different-source data. The smaller this empirical probability, the less likely it is that the pair (A^*, B^*) was generated by different sources.

This approach is similar in spirit to the use of simulation envelopes for computing confidence intervals for a spatial assocation function in spatial point process models (e.g. Baddeley *et al.* (2014)) where resampling techniques are used to estimate confidence intervals under a null model (such as complete spatial randomness). In our approach the null model assumes that the two event series were generated by different sources, as discussed in the next section.

5.2.2. Sessionized resampling

For our different-source hypothesis we use a null model that assumes that A^* is conditionally independent of B^* given an inhomogeneous background intensity process (e.g. that varies with the time of day for user activity). In particular, we generate simulated series A' that depend on the background intensity and that have similar marginal characteristics to the observed series A^*. The particular details of how the simulation is carried out can be domain specific. Since the user-generated event data that are of interest to this paper are typically inhomogeneous and bursty, we pursue an approach that we call *sessionized resampling*.

Specifically, we keep the event times in B^* fixed and generate multiple random realizations A' of A^* by randomly perturbing the event times in A^*. In particular, to preserve the bursty and inhomogeneous nature of the data, we work with sessions (collections of event times) rather than individual event times. Sessions are defined formally below. We sample new times for the starts of sessions (rather than new times for individual events). Each session is then shifted to the corresponding sampled (perturbed) start time. The new session start times are sampled from an inhomogeneous background distribution over time. We next describe the steps in this approach in more detail.

To sessionize the data we proceed by defining the first event in a session to be any event that occurs after a period of T or more time units of inactivity (for example, see Spiliopoulou *et al.* (2003)). We define the set of session start times for series A^* as

$$\begin{aligned} A^*_\mathrm{ses} &= \{t_j : j = 1 \text{ or } t_j - t_{j-1} \geqslant T \text{ for } j = 2, \ldots, n_{A^*}\} \\ &\equiv \{t_{\mathrm{ses},k} : k = 1, \ldots, r_{A^*}\}. \end{aligned} \tag{10}$$

Thus A^* has $r_{A^*} \leqslant n_{A^*}$ sessions, where a session is defined as all of the events after one session start and before the next. We then define the sessionized series A^* to be composed of the process of session start times A^*_ses and the event times in each corresponding session. Namely, $A^* = \cup_{k=1}^{r_{A^*}} S_k$ where

Table 1. Algorithm 1: sessionized resampling

Input: pair of event series (A, B); resampling time distribution $p(t_{\text{ses}})$
Output: set of n_{sim} resampled pairs $\mathcal{D}$
Fix B (where B is the shorter series, i.e. $n_B < n_A$)
Derive the r_A sessions $S_k, k = 1, \ldots, r_A$, of A as defined in the text
for $l = 1$ to n_{sim} *do*
  for $k = 1$ to r_A *do*
    draw $t_{\text{new}} \sim p(t_{\text{ses}})$
    elementwise, set $S_k^{(l)} = S_k - t_{\text{ses},k} + t_{\text{new}}$
  end for
  Set $A^{(l)} = \cup_{k=1}^{r_A} S_k^{(l)}$
end for
return $\mathcal{D} = \{(A^{(l)}, B) : l = 1, \ldots, n_{\text{sim}}\}$

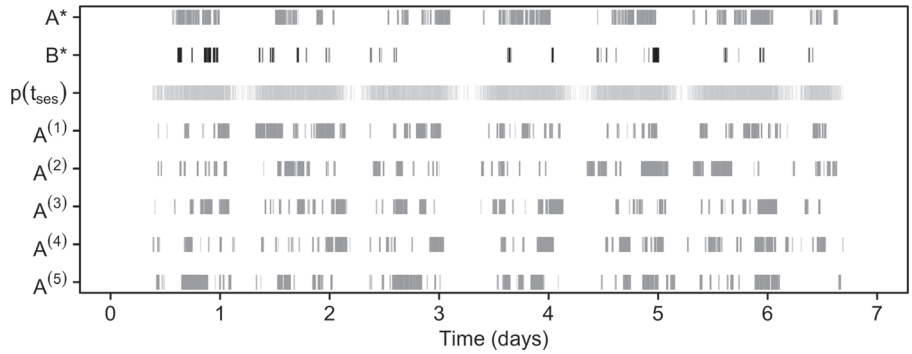

Fig. 3. Example of sessionized resampling for a pair of event series (A^*, B^*) taken from the student web browsing data: here we use $T = 10$ min, and the distribution of session start times $p(t_{\text{ses}})$ is the empirical distribution of session start times across all series A available in the data set; $A^{(l)}$ for $l = 1, \ldots, 5$ represent five event series simulated via algorithm 1

$$S_k = \begin{cases} \{t_j : t_{\text{ses},k} \leqslant t_j < t_{\text{ses},k+1}\} & \text{if } k = 1, \ldots, r_{A^*} - 1, \\ \{t_j : t_{\text{ses},k} \leqslant t_j \leqslant t_{n_{A^*}}\} & \text{if } k = r_{A^*}. \end{cases} \tag{11}$$

This definition leaves the event series unchanged but groups activity according to bursts of activity.

A replicate of A^* can be generated as follows. Sample r_{A^*} new session start times from a distribution of session start times $p(t_{\text{ses}})$. (In general any distributional form for $p(t_{\text{ses}})$ that reflects the inhomogeneous nature of event series could be used.) Then shift all events in each of the sessions S_k for $k = 1, \ldots, r_{A^*}$ so that the first event $t_{\text{ses},k}$ occurs at the kth newly sampled session start time. This process preserves the total number of events in A^* as well as the number of and spacing of events in each session. See algorithm 1 in Table 1 for the pseudocode to generate resampled series and Fig. 3 for an illustration of how the approach works.

6. Results

Below we describe experimental results for

(a) a simulation study of homogeneous event time series and
(b) two case-studies of real user event data.

6.1. Simulation

We simulated event series to form pairs of temporal processes, both independent pairs and pairs with varying degrees of association, to assess the behaviour of our proposed methods for computing score-based likelihood ratios and CMPs. The simulated series have similar marginal characteristics (in terms of overall rates of event generation) to the data from our first case-study (the student web browsing data that are described in Section 6.2) where individuals generate two types of events, at different rates per individual, over a period of 1 week.

The simulation process was as follows. We generate A events over a window of 1 week from a Poisson process with rate λ_A, where λ_A is sampled from a kernel density estimate of observed event rates across different users from the case-study with web browsing data. The rate of events for series B is proportional to λ_A, with $\lambda_B = p\lambda_A$ where $p \in [0, 1]$ is the relative frequency of type B events to type A events.

For independent series we simulate events independently from two Poisson processes with rates λ_A and λ_B. For dependent processes we again simulate A-events from a Poisson process with rate λ_A, but now generate B events according to algorithm 2 in Table 2. For every simulated A-event, a B-event is generated with probability p, and (if generated) the time of the B-event is distributed via a Gaussian density with standard deviation σ centred at the time of the A-event. (We also experimented with sampling from non-Gaussian distributions such as the exponential distribution and obtained similar results to those described here.) The degree of association between two simulated processes is controlled by

(a) the relative frequency p of events of type B to events of type A,
(b) the standard deviation σ and
(c) the intensity λ_A of process A which also controls the number of events in both A and B.

Our ability to detect an association is expected to decrease as

(i) p decreases and
(ii) σ increases.

The relationship between detectability and the number of events in each process is more complex, as we discuss later.

Simulations were performed with different combinations of parameter settings to investigate the sensitivity of our detection methods across a variety of scenarios. To ensure sufficient variation in the event counts we sampled the rates for process A (λ_A) in algorithm 2 from a kernel density estimate as described earlier and multiplied the sampled intensity by a rate multiplier

Table 2. Algorithm 2: simulation of associated marked point processes

Input: intensity λ_A, relative frequency of B events to A events p, standard deviation σ
Output: simulated pair of processes (A, B)
Simulate $A = \{t_j : j = 1, \ldots, n_A\}$ from a Poisson point process with rate λ_A
Set $k = 0$
for $j = 1$ to n_A *do*
  draw $d_j \sim$ Bernoulli(p)
  if $d_j = 1$ *then*
    increment $k = k + 1$
  draw $t_k \sim N(\mu = t_j, \sigma^2)$
  end if
end for
return $A = \{t_j : j = 1, \ldots, n_A\}$, $B = \{t_k : k = 1, \ldots, n_B = \Sigma_{j=1}^{n_A} d_j\}$

$r \in \{1, 10\}$ to assess the effect of dense event series on the SLR and CMP. For a given setting of parameter values (r, p, σ), we simulated both independent and dependent series. The relative frequency of type B events to type A events was one of $p \in \{0.01, 0.10, 0.20, 0.50, 0.75, 0.95\}$. Finally, the standard deviation of the Gaussian distribution that was used to generate time stamps for events of type B was $\sigma \in \{0.5, 1, 2, 5, 10\}$ min.

For each combination of parameters (r, p, σ), we generated 10000 independent event series pairs and 10000 event series pairs with association and computed SLR and CMP values for each pair. For the SLR, we utilized the leave-pairs-out cross-validation methodology that was described earlier in Section 5.1. We then thresholded the ranked scores to obtain binary decisions and compared the binary decisions with the known ground truth (independent pairs are treated as different source and associated pairs (regardless of the strength of association) are treated as same source) from the simulation to compute both true and false positive rates. Further, we varied the threshold to achieve different trade-offs in terms of sensitivity and specificity. The area under the receiver operating characteristic curve, AUC, can be used to summarize this trade-off. AUC is a measure of goodness of fit and can be thought of as the probability that the method will result in a larger SLR (or smaller CMP) for a randomly chosen same-source pair than that of a randomly chosen different-source pair (Fawcett, 2006; Krzanowski and Hand, 2009).

In general we found that we could detect associated event series pairs over a wide variety of parameter settings, for both the segregation index and interevent time statistics and for both SLR and CMP methods. Here we focus on results for the mean interevent time score function (the segregation score was not as accurate in detection). We found that the two most important factors in assessing the performance of our methods are the number of events in process B and the SNR, defined as

$$\mathrm{SNR} \equiv (\overline{\lambda}_A)^{-1}/\sigma. \tag{12}$$

Here, the 'signal' is inversely proportional to the standard deviation σ of the Gaussian distribution that is used to generate event times in B. Smaller values of σ correspond to smaller interevent times from events in B to events in A and, therefore, higher signal. The 'noise' is the reciprocal of the mean intensity across the simulated realizations of process A, denoted $(\overline{\lambda}_A)^{-1}$. As this value decreases, the noise (or the density of events in realizations of A) increases. As an extreme case, consider a single process A' with $\lambda_{A'}^{-1} \to \infty$, which implies that $\mathbb{E}(\overline{\mathcal{T}}_{A'A'}) \to 0$. Regardless of the strength of the signal, events in B will occur close in time to events in A', and therefore any other series will appear to be associated with A'. The SNR controls for this phenomenon. As the SNR increases we expect the association of two event series to be more easily detected via methods such as the SLR or CMP. For the simulation study, the SNR is known. In practice, a natural estimator of the SNR for a single pair of event series (A, B) is given by $\widehat{\mathrm{SNR}} = \overline{\mathcal{T}}_{AA}/\overline{\mathcal{T}}_{BA}$.

Fig. 4 shows the detectability of association of simulated event series via the SLR (Figs 4(a), 4(b), 4(d) and 4(e)) and CMP (Figs 4(c) and 4(f)) as a function of the SNR of the simulated series. (We considered five values of σ and two values of r, but when computing the SNR there are only eight unique values due to the overlap of $\sigma \in \{5, 10\}$ with $r = 1$ and $\sigma \in \{0.5, 1\}$ with $r = 10$, e.g. $(1\overline{\lambda}_A 5)^{-1} = (10\overline{\lambda}_A 0.5)^{-1}$. Furthermore, $(\overline{\lambda}_A)^{-1} = 7.3$ min, which is the corresponding mean interevent time in the first case-study.) Figs 4(a)–4(c) show boxplots of the values of SLR and CMP, and the points in Figs 4(d)–4(f) show the corresponding AUC-values, each as a function of the SNR. Here we present results for simulations with a relative frequency of $p = 0.2$. Results for other values of p are qualitatively similar (the magnitude of the SLR increases slightly for small SNR as p increases, but the CMPs are indistinguishable for varying p). As the SNR

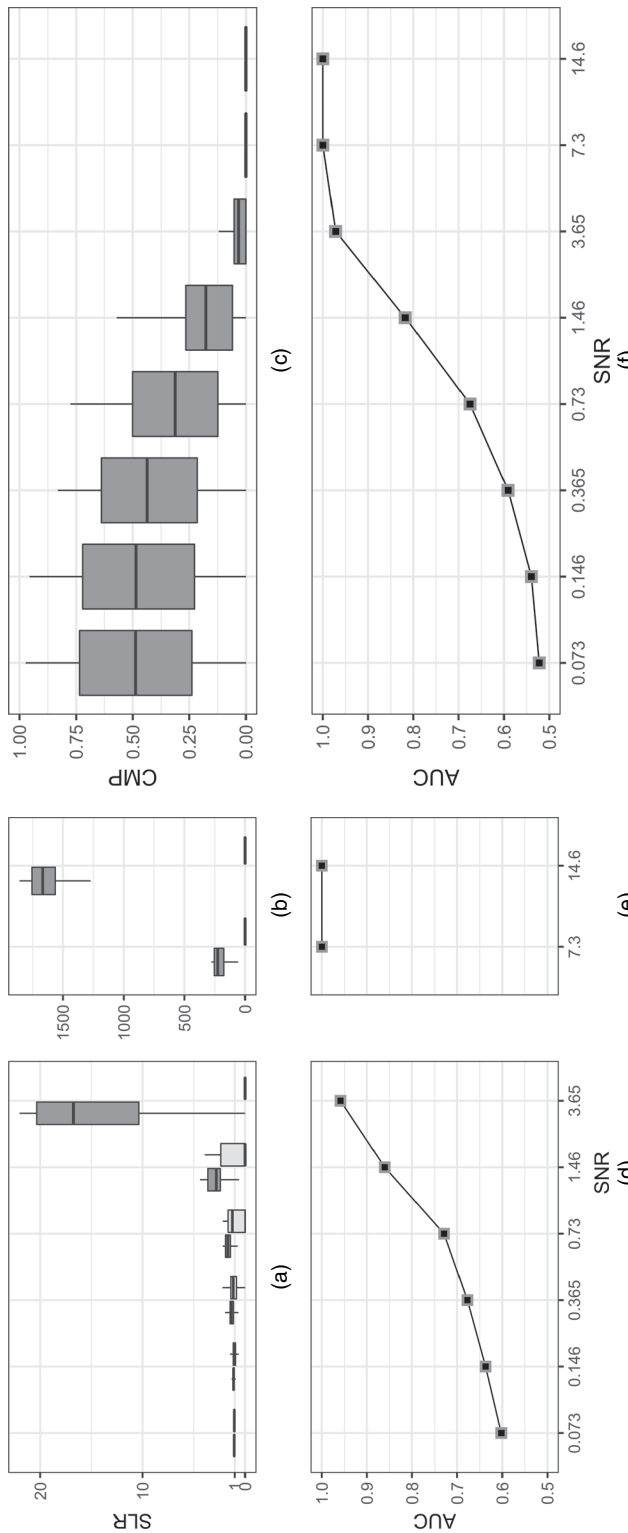

Fig. 4. (a)–(c) Boxplots of measures of association for $\Delta(A,B) = \overline{T}_{BA}$ with $p = 0.20$ and (d)–(f) corresponding AUC-values as a function of the SNR: (a), (b), (d), (e) score-based likelihood ratio (note the different scales on the *y*-axis in (b)); (c), (f) CMP (note that the CMPs for independent pairs are uniformly distributed by definition and thus have been omitted from (c)); ▢, independent event series pairs; ▢, associated event series pairs

increases, the SLR of associated pairs increases, the CMP decreases and AUC increases, all of which are indicative of being better able to separate associated and independent pairs of event series. Note that as the SNR grows large (e.g. SNR > 7.3) the AUC-values of both classifiers have an asymptote near 1 and CMP values near 0 for associated pairs, but the SLR of associated pairs continues to increase. This implies that both methods perform similarly for classification, but that the SLR is better calibrated with values increasing indefinitely as the SNR increases.

Fig. 5 overlays the AUC-curves for SLR and CMP from Fig. 4 on the same plot. The SLR performs better than the CMP for low values of the SNR, indicating that the same source score distribution aids in quantification of degree of association when the SNR is low, but both techniques perform similarly when SNR $\geqslant 3.65$. Therefore, the CMP results in no information loss compared with the SLR for pairs of processes exhibiting high degrees of association.

In addition to the SNR, the number of events in series B also influenced the detectability of association. We found little sensitivity in the SLR to p for any given intensity λ_A, but varying both together resulted in dramatically different behaviour. We use n_B as a proxy for the combination of relative frequency and rate because it is a stochastic function of the two such that $\mathbb{E}(n_B) = \omega p \lambda_A$ where ω is the length of the observation window. Fig. 6 depicts a non-parametric regression of SLR on n_B for associated pairs of processes with $p = 0.20$. If the observed number of events is small, then the score function (the mean interevent time) will have higher variance. The high

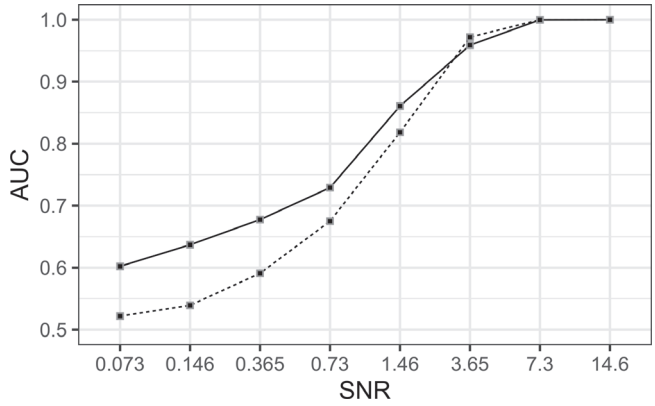

Fig. 5. AUC-values for both the SLR (———) and the CMP (--------) as a function of the SNR for simulated data with $p = 0.20$

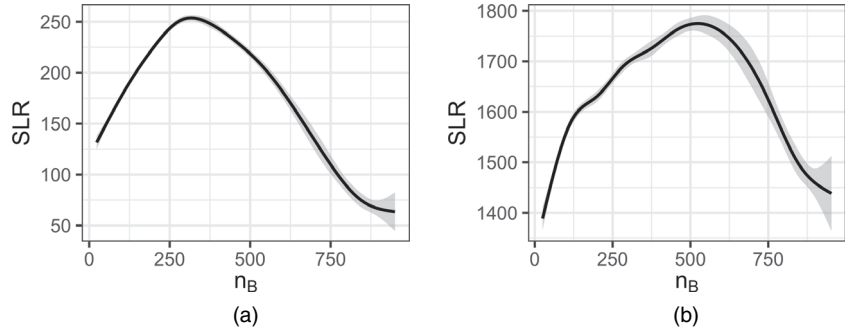

Fig. 6. Generalized additive model smoother of the SLR for simulated associated pairs with $p = 0.20$ as a function of the number of events in series B (———, smoother fit; ▬, 99% confidence interval): (a) SNR = 7.3; (b) SNR = 14.6 (note the different scales on the y-axis)

variance in the score function would be expected under both the same-source and the different-source distributions, generally leading to smaller SLR values. Conversely, if this number is large the processes become so dense that the interevent times for independent pairs decrease (i.e. interevent times under the different source hypothesis) and, therefore, behave more like the interevent times of associated pairs. In either extreme, we observe lower SLRs relative to the peak value that occurs in the middle of this range. For these data the nature of the conclusion does not change (i.e. the SLR favours the same-source hypothesis across the entire range of n_B-values) but the observed patterns suggest relationships that may be important with other data.

Note that the trend that is exhibited in Fig. 6 is consistent across all values of the relative frequency p and SNR. However, the magnitudes differ with smaller SNR yielding a lower peak SLR. This makes sense intuitively because, as the SNR decreases, the nearest neighbours of events in B are no longer the generating events in A (i.e. they can be closer to another unassociated event). (Here 'generating event' refers to the time t_j of the simulated event in A which was used as the mean of the Gaussian distribution to generate a given event in B in algorithm 2.) Thus, process B behaves similarly to an independently generated Poisson process when the SNR is small.

The results of the simulation study illustrate the promise of the resampling approach for calculating CMPs in situations where no reference data are available. The population-based SLR is still the preferred method, however, given its better performance for pairs exhibiting weak association and the fact that it performs similarly to the CMP for strongly associated pairs.

6.2. Case-study I—student web browsing data

The data that are considered in this section come from an *in situ* observational study of student activity over time on digital devices, conducted at a large US university (Wang *et al.*, 2015). 124 undergraduate students with Windows computers voluntarily participated in the study for 1 week and browser activity was automatically logged. Participants were instructed to continue using their devices as normal.

The event logs were dichotomized by the type of web browsing event to create pairs of event time series (A, B) for each student. Series B corresponds to Facebook events (i.e. any web browser activity occurring on facebook.com), and series A corresponds to non-Facebook events (i.e. any web browser activity not occurring on facebook.com). Students were included in our analysis if they had at least 50 events of each type. Of the 124 students who were originally recorded, 55 met the inclusion criteria. These students generated 90340 log records, with 13995 (15.5%) Facebook and 76345 (84.5%) non-Facebook browser events. A graphical illustration of a subset of the data is shown in Fig. 7.

6.2.1. Population-based results

Fig. 8 shows the empirical distributions of each of the score functions for same- and different-source pairs as discussed in Section 4. Note that all pairwise combinations of the data were included in the reference data set $\mathcal{D}$ that was used to create these densities for illustration (leave-pairs-out cross-validation was used for the rest of the results in this section). Although there is some overlap in the same- and different-source densities for all score functions, it is clear that the majority of the probability mass does not occur in the same region. This suggests that both the SLR and the CMP should be able to assess the strength of association accurately.

Using the SLRs that were estimated via leave-pairs-out cross-validation and a threshold of 1, which corresponds to the data being equally likely to have been generated under either hypothesis, we can compare the true and false positive rates for each score function. Table 3

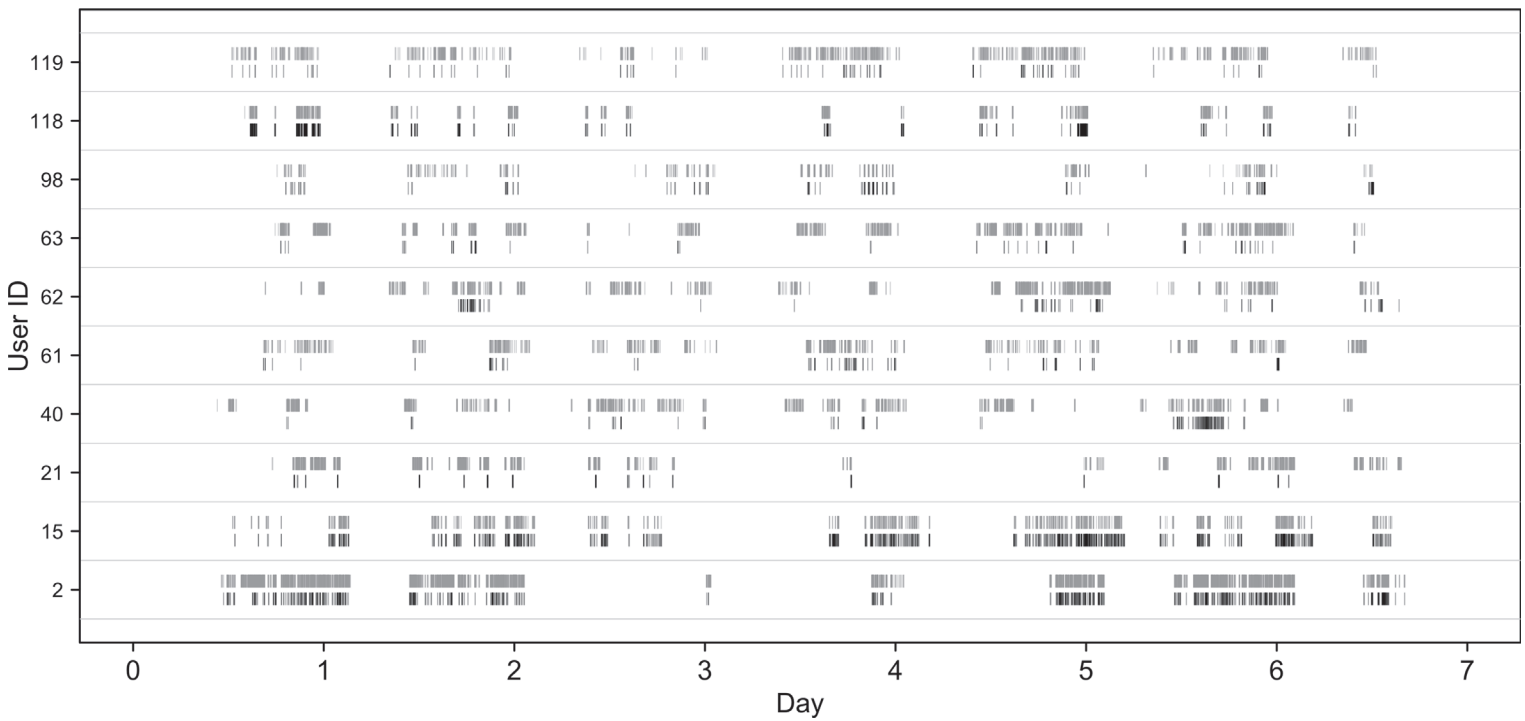

Fig. 7. Web browsing data observed over 7 days from a random sample of 10 users from the case-study data: each user has two rows corresponding to the two event series with the top row of grey bars of non-Facebook web browsing events (A_i) and the bottom row of black bars of Facebook events (B_i); note that all events shown above are relative to the first day of observation for each student, and each tick mark on the x-axis represents midnight of the corresponding day

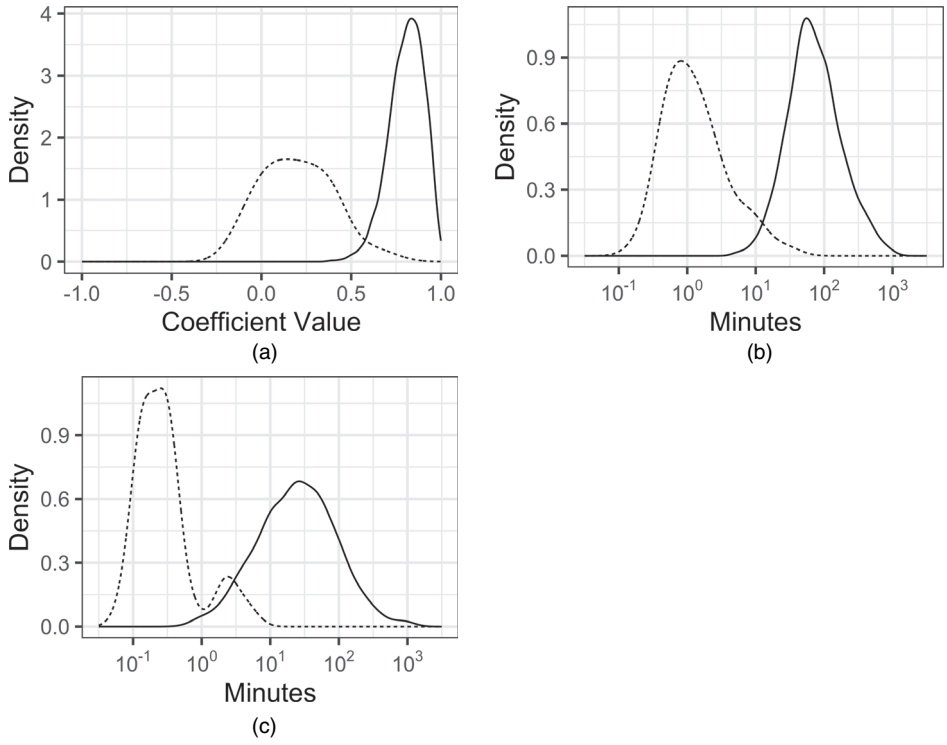

Fig. 8. Empirical distributions of the score functions from Section 4 (same-source distributions (H_s, - - - - -) and different-source distributions (H_d, ———) approximated via kernel density estimation with Gaussian kernels and Scott's rule-of-thumb bandwidth; leave-pairs-out cross-validation was not used to produce these densities; instead all pairs were used for illustration: (a) segregation; (b) mean interevent time; (c) median interevent time

provides these rates (listed as TP@1 and FP@1 respectively). Note that the SLR based on the mean interevent time yields the highest true positive rate and lowest false positive rate for this particular choice of threshold. The area under the receiver operating characteristic curve is also given in Table 3. The choice of score function seems to have little effect on AUC, since all score functions yield a value that is greater that 0.99.

Perhaps of most interest to forensic scientists is the threshold of SLR values that gives a 0% false positive rate, since wrongfully convicting the innocent has extremely negative societal consequences. Table 3 lists this threshold and its corresponding true positive rate (TP@FP=0) for all the score functions that were considered. The interevent time summary statistics require a lower SLR than the marked point process characteristics and yield a higher true positive rate, which is evidence that they are better calibrated. Note that a pair of event series (A, B) whose SLR value is equal to 122 for the mean interevent time $\overline{T}_{BA}$ has the following interpretation: the observed mean interevent time for pair (A, B) was 122 times more likely to have been generated by same-source series than by different-source series.

6.2.2. Resampling results

We now consider the case where we have only a single pair of event series (A^*, B^*) available for analysis—for example, for any pair of event series in our case-study data we would like to assess the degree of association between that pair only by using the data for the pair. We

Table 3. Performance of a classifier based on SLR_{Δ}

Score function	*TP@1*	*FP@1*	*FP = 0 threshold*	*TP@FP = 0*	*AUC*
S	0.945	0.022	1871	0.745	*0.992*
$\overline{\mathcal{T}}_{BA}$	*0.964*	*0.021*	*122*	*0.873*	*0.992*
$\mathrm{med}(\mathcal{T}_{BA})$	0.945	0.06	279	0.818	0.99

Table 4. Performance of a classifier based on CMP_{Δ}

Score function	*TP@.05*	*FP@.05*	*TP@.001*	*FP@.001*	*AUC*
$\overline{\mathcal{T}}_{BA}$	*1.000*	*0.036*	0.982	*0.002*	*0.999*
$\mathrm{med}(\mathcal{T}_{BA})$	*1.000*	0.176	*1.000*	0.015	0.992

used sessionized resampling with B^* fixed to estimate the CMP for each pair of series by using both the mean and the median interevent times. To compare with the results of the population-based approach more directly, $p(t_{\mathrm{ses}})$ was chosen to be the empirical distribution of all session start times in the data excluding those from the particular pair (A^*, B^*) being analysed in a fashion similar to leave-pairs-out cross-validation. We define the estimated CMP as the fraction of simulated pairs whose score function is less than that of the observed pair.

For each of the 55 same-source pairs 10000 samples were generated via sessionized resampling, whereas, for each of the 2970 different-source pairs, 1000 samples were generated (resulting in approximately 3 million total iterations of the sampler). Similarly to our population-based approach, we can view the CMP as a discriminant function for a binary classification decision (for example, pairs with CMP values that are less than some threshold are considered same source) and compare the true and false positive rates for each score function. Table 4 provides these rates for thresholds of 0.05 and 0.001. Note that a pair of event series (A^*, B^*) whose CMP value is equal to one of these thresholds has the following interpretation: the probability that a score that is comparable with or lower than that obtained from the pair (A^*, B^*) was generated by different-source event series is 5% (or 0.1%).

The CMP worked quite well in quantifying the degree of association for these data, as evidenced by the AUC-values in Table 4. In fact the CMP performed better than the SLR when using AUC as the evaluation metric. We also observed this phenomenon in our simulation study for pairs of processes exhibiting a high SNR, which implies that this particular data set is well suited for the resampling approach because it is comprised of highly associated pairs of same-source event series.

6.3. Case study II—Los Alamos National Laboratory authentication data

Suppose that an examiner is given two sets of time-stamped authentication, or log-in, events. The two series of events are composed of log-ins to a user's personal computer and a shared computer. The examiner is tasked with quantifying the association between these event series to determine whether both were generated by the same user (i.e. the user whose personal computer authentication event series was collected) or not. Further assume that the examiner is only

presented with the event series in question and does not have access to a population of similar event series.

We show a proof of concept of the efficacy of the CMP on this task with real authentication data. The data represent successful authentication events from users to computers on the Los Alamos National Laboratory enterprise network (Kent, 2014). Each authentication event is composed of its time stamp (represented by the number of seconds from some unknown origin time), the user account that generated the event (the *actor*) and the computer that the actor logged into (the *target*). Note that the actors and targets were anonymized. We focus on two users—U4116 and U7250—and their authentication events to three particular computers—C248, C4751 and C8268—during the first day of available data. Both users authenticate to computer C248, and computers C4751 and C8268 are logged into only by users U4116 and U7250 respectively, and not by any other users in the entirety of the authentication data. Table 5 gives the number of authentication events for each user–computer pair, and Fig. 9 shows the event series themselves.

We computed the CMP for each pairwise combination of event series available. Thus, there were two same-source pairs (user U4116's authentications to machines C248 and C4751, denoted U4116–C248 and U4116–C4751; and user U7250's authentications to machines C248 and C8268, denoted U7250–C248 and U7250–C8268) and two different-source pairs (U4116–C248 and U7250–C8268, and U7250–C248 and U4116–C4751). In each case, authentication events to the shared machine (series B) were fixed and the event times of authentications to the unique machine (series A) were resampled 10000 times via sessionized resampling. Session start

Table 5. Number of log-in events for each user to each computer on the first day of activity in the Los Alamos National Laboratory authentication data

User	*Results for the following computers:*		
	C248	*C4751*	*C8268*
U4116	120	318	0
U7250	53	0	362

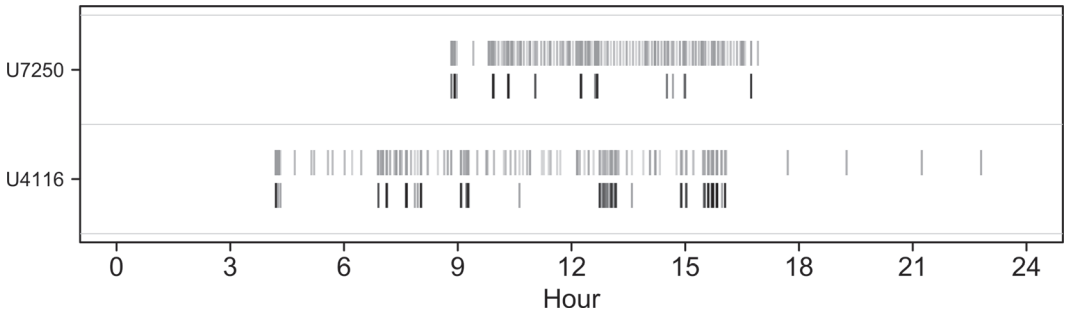

Fig. 9. Los Alamos National Laboratory authentication data: shared machine (|) refers to computer C248, and unique machine (|) refers to computers C4751 and C8268 for users U4116 and U7250 respectively

Table 6. CMPs for various score functions for the Los Alamos National Laboratory authentication data†

Unique machine	*Shared machine*	*Mean interevent time*	*Median interevent time*	*Segregation*
U4116–C4751	U4116–C248	0.000	0.000	0.000
U7250–C8268	U7250–C248	0.000	0.000	0.000
U4116–C4751	U7250–C248	0.197	0.637	0.952
U7250–C8268	U4116–C248	0.545	0.266	0.333

†Lower scores are indicative of same-source event series.

times were sampled from a Gaussian distribution with mean equal to the centre of the fixed event series (authentications to the shared computer) and with a standard deviation such that 99% of the session starts fall in the range of that fixed series. Note that other distributions over session start times, including uniform and empirical distributions, yield similar conclusions. The resulting CMPs for each score function discussed in Section 4 are provided in Table 6.

Across all score functions, the same-source authentication event series for both users U4116 and U7250 exhibit CMPs equal to 0, which is strongly indicative that they were in fact generated by the same user. Conversely, the different-source event series exhibit CMPs ranging from 0.20 to 0.95, indicating that the association of these series was at a level that would be typical for different-source series and unlikely for same-source series. Overall, the resampling-based approach proved effective for this particular data set.

7. Conclusion

Drawing on previous work from the forensics and statistics literature, we explored a variety of measures for quantifying the association between two discrete event time series. Multiple score functions were used to determine the similarity between the series, including characteristics of marked point processes (the coefficient of segregation) and interevent time summary statistics (the mean and median). These score functions were shown to be discriminative for same- and different-source pairs of event series.

We then proposed two methods for assessing the strength of association for a given pair of event series. The population-based approach uses a sample from the relevant population to construct SLRs that assess the relative likelihood of observing a given degree of association when the series came from the same or different sources. The resampling approach considers only a single pair of event series, simulates a different-source score distribution via sessionized resampling and uses that distribution to calculate CMPs.

Although the population-based approach with SLRs remains the preferred technique in terms of accuracy and interpretability, our proposed resampling technique with CMPs shows considerable promise for assessing the degree of association of pairs of user-generated event series. However, both techniques require more extensive study and testing before being used in practice by forensic examiners.

Future directions include investigating different types of association, incorporating more information in the marked point processes and developing realtime anomaly detection algorithms. In this paper, we focused on detecting the interleaving of bursty, inhomogeneous processes, but other types of dependence (e.g. lagged or 'triggered' bursts of events of different type) warrant future study. One could also include spatial information in the marked point process from the

Global Positioning System of devices or consider more than two types of events (e.g. geolocated smartphone data with marks corresponding to actions across different applications, short message service and e-mails). This additional information could result in higher accuracy for both methods. In the case of multiple-event series, the techniques could also be extended for use in pattern mining to determine which event series are associated with one another. Techniques for overcoming multiple-testing complications would need to be developed under this scenario. Overall, work in the area of quantifying association of user event data shows considerable promise and potential efficacy in both forensic and cybersecurity settings.

Acknowledgements

The authors thank Gloria Mark at the University of California, Irvine, for providing the web browsing data that were used in this paper. We also thank Juston Moore from Los Alamos National Laboratory for useful discussions regarding the authentication data.

This research was partially funded through co-operative agreement #70NANB15H176 between the US National Institute of Standards and Technology and Iowa State University, which includes activities carried out at Carnegie Mellon University, the University of California, Irvine, and University of Virginia. Author PS was also funded in part by US National Science Foundation award IIS-1320527.

References

Aitken, C. and Stoney, D. (1991) *The Use of Statistics in Forensic Science*. Chichester: Horwood.

Årnes, A. (2017) *Digital Forensics*. Hoboken: Wiley.

Baddeley, A., Diggle, P. J., Hardegen, A., Lawrence, T., Milne, R. K. and Nair, G. (2014) On tests of spatial pattern based on simulation envelopes. *Ecol. Monogr.*, **84**, 477–489.

Baddeley, A., Rubak, E. and Turner, R. (2015) *Spatial Point Patterns: Methodology and Applications with R*. Boca Raton: Chapman and Hall–CRC.

Boers, N., Bookhagen, B., Marwan, N. and Kurths, J. (2016) Spatiotemporal characteristics and synchronization of extreme rainfall in South America with focus on the Andes mountain range. *Clim. Dynam.*, **46**, 601–617.

Bolck, A., Ni, H. and Lopatka, M. (2015) Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison. *Law, Probab. Risk*, **14**, 243–266.

Bozza, S., Taroni, F., Marquis, R. and Schmittbuhl, M. (2008) Probabilistic evaluation of handwriting evidence: likelihood ratio for authorship. *Appl. Statist.*, **57**, 329–341.

Casey, E. (2011) *Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet*. New York: Academic Press.

Champod, C. and Evett, I. W. (2001) A probabilistic approach to fingerprint evidence. *J. Forens. Identificn*, **51**, 101–122.

Champod, C. and Meuwly, D. (2000) The inference of identity in forensic speaker recognition. *Speech Commun.*, **31**, 193–203.

Daley, D. J. and Vere-Jones, D. (2003) *An Introduction to the Theory of Point Processes*, vol. II, *General Theory and Structure*, 2nd edn. New York: Springer.

Diggle, P. J. and Chetwynd, A. G. (1991) Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, **84**, 1155–1163.

Dixon, P. M. (2014) Ripley's *k* function. *Wiley StatsRef: Statistics Reference Online*.

Donges, J. F., Schleussner, C.-F., Siegmund, J. F. and Donner, R. V. (2016) Event coincidence analysis for quantifying statistical interrelationships between event time series. *Eur. Phys. J. Specl Top.*, **225**, 471–487.

Evett, I. W. and Weir, B. S. (1998) Presenting evidence. In *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists* (eds J. Fagerberg, D. C. Mowery and R. R. Nelson), ch. 9, pp. 235–266. Sunderland: Sinauer.

Fawcett, T. (2006) An introduction to ROC analysis. *Pattn Recogn. Lett.*, **27**, 861–874.

Gaines, K. F., Bryan, Jr, A. L. and Dixon, P. M. (2000) The effects of drought on foraging habitat selection of breeding wood storks in coastal Georgia. *Waterbirds*, **23**, 64–73.

Galbraith, C. and Smyth, P. (2017) Analyzing user-event data using score-based likelihood ratios with marked point processes. *Digtl Investign*, **22**, specl iss., S106–S114.

Hanisch, K. (1984) Some remarks on estimators of the distribution function of nearest neighbour distance in stationary spatial point processes. *Ser. Statist.*, **15**, 409–412.

Hepler, A. B., Saunders, C. P., Davis, L. J. and Buscaglia, J. (2012) Score-based likelihood ratios for handwriting evidence. *Forens. Sci. Int.*, **219**, 129–140.

Hunter, J. D. and Milton, J. G. (2003) Amplitude and frequency dependence of spike timing: implications for dynamic regulation. *J. Neurphysiol.*, **90**, 387–394.

Illian, J., Penttinen, A., Stoyan, H. and Stoyan, D. (2008) *Statistical Analysis and Modelling of Spatial Point Patterns*. Chichester: Wiley.

Kent, A. D. (2014) User-computer authentication associations in time. Los Alamos National Laboratory, Los Alamos.

Krzanowski, W. J. and Hand, D. J. (2009) *ROC Curves for Continuous Data*. Boca Raton: Chapman and Hall–CRC.

Loh, J. M. (2008) A valid and fast spatial bootstrap for correlation functions. *Astrophys. J.*, **681**, 726–734.

Malik, N., Marwan, N. and Kurths, J. (2010) Spatial structures and directionalities in monsoonal precipitation over South Asia. *Nonlin. Process. Geophys.*, **17**, 371–381.

Meuwly, D., Ramos, D. and Haraksim, R. (2017) A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forens. Sci. Int.*, **276**, 142–153.

Myers, S. P., Timken, M. D., Piucci, M. L., Sims, G. A., Greenwald, M. A., Weigand, J. J., Konzak, K. C. and Buoncristiani, M. R. (2011) Searching for first-degree familial relationships in California's offender DNA database: validation of a likelihood ratio-based approach. *Forens. Sci. Int. Genet.*, **5**, 493–500.

Niehof, J. T. and Morley, S. K. (2012) Determining the significance of associations between two series of discrete events: bootstrap methods. *Report LA-14453*. Los Alamos National Laboratory, Los Alamos.

Oh, J., Lee, S. and Lee, S. (2011) Advanced evidence collection and analysis of web browser activity. *Digitl Investign*, **8**, specl iss., S62–S70.

Pielou, E. (1977) *Mathematical Ecology*. New York: Wiley.

Quiroga, R. Q., Kraskov, A., Kreuz, T. and Grassberger, P. (2002) Performance of different synchronization measures in real data: a case study on electroencephalographic signals. *Phys. Rev.* E, **65**, article 041903.

Radicchi, F. (2009) Human activity in the web. *Phys. Rev.* E, **80**, article 026118.

Ross, A., Nandakumar, K. and Jain, A. (2006) *Handbook of Multibiometrics*. New York: Springer.

Roussev, V. (2016) Digital forensic science: issues, methods, and challenges. In *Synthesis Lectures on Information Security, Privacy, and Trust* (eds E. Bertino and R. Sandhu). San Rafael: Morgan and Claypool.

Roussev, V. and McCulley, S. (2016) Forensic analysis of cloud-native artifacts. *Digitl Investign*, **16**, specl iss., S104–S113.

Scott, D.W. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley.

Siegmund, J. F., Siegmund, N. and Donner, R. V. (2017) Coincalc—a new *R* package for quantifying simultaneities of event series. *Comput. Geosci.*, **98**, 64–72.

Spiliopoulou, M., Mobasher, B., Berendt, B. and Nakagawa, M. (2003) A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *INFORMS J. Comput.*, **15**, 171–190.

Stern, H. S. (2017) Statistical issues in forensic science. *A. Rev. Statist. Appl.*, **4**, 225–244.

Synowiecki, R. (2007) Consistency and application of moving block bootstrap for non-stationary time series with periodic and almost periodic structure. *Bernoulli*, **13**, 1151–1178.

Thompson, W. C. and Newman, E. J. (2015) Lay understanding of forensic statistics: evaluation of random match probabilities, likelihood ratios, and verbal equivalents. *Law Hum. Behav.*, **39**, 332–349.

Wang, Y., Niiya, M., Mark, G., Reich, S. and Warschauer, M. (2015) Coming of age (digitally): an ecological view of social media use among college students. In *Proc. 18th Conf. Computer Supported Cooperative Work and Social Computing*, pp. 571–582. New York: Association for Computing Machinery.

Willis, S., McKenna, L., McDermott, S., O'Donell, G., Barrett, A., Rasmusson, B., Nordgaard, A., Berger, C., Sjerps, M., Molina, J. J. L., Zadora, G., Aitken, C., Lunt, L., Champod, C., Biedermann, A., Hicks, T. N. and Taroni, F. (2016) ENFSI guideline for evaluative reporting in forensic science. European Network of Forensic Science Institutes, Wiesbaden. (Available from `http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf`.)